Limit theorems for sequences of random trees

David Balding, Pablo A. Ferrari, Ricardo Fraiman*
and Mariela Sued

Department of Epidemiology and Public Health, Imperial College, England
Instituto de Matemática e Estatística, Univ. de São Paulo, Brasil
Departamento de Matemática y Ciencias, Univ. de San Andrés, Argentina
and Centro de Matemática, Univ. de la República, Uruguay
Instituto del Cálculo, Univ. de Buenos Aires, Argentina

February 1, 2008

Running Title: Limit theorems on random trees

Key words: random trees, d-mean, invariance principle, Kolmogorov-Smirnov goodness-of-fit test.

AMS Subject Classification: Primary: 60F17, 60D05.

Abstract

We consider a random tree and introduce a metric in the space of trees to define the "mean tree" as the tree minimizing the average distance to the random tree. When the resulting metric space is compact we have laws of large numbers and central limit theorems for sequence of independent identically distributed random trees. As application we propose tests to check if two samples of random trees have the same law.

1 Introduction

Random trees have long been an important modelling tool. In particular, trees are useful when a collection of observed objects are all descended from a common ancestral object

*Corresponding author. Postal address: Departamento de Matemática y Ciencias, Universidad de San Andrés, Vito Dumas 284, 1644, Victoria, Argentina. *Tel.*: 54-11-47257062. *Fax*: 54-11-47257010. Email address: `rfraiman@udesa.edu.ar`

via a process of duplication followed by gradual differentiation. This characterizes the process of natural evolution, and also any form of information that over time is successively replicated, and transmitted with occasional error. There are two broad approaches to constructing random evolutionary trees: forwards in time "branching process" models, such as the Galton-Watson process, and backwards-in-time "coalescent" models such as Kingman's coalescent (Kingman, 1982).

We prove law of large numbers and an invariance principle for random trees defined in a metric space and propose a Kolmogorov-Smirnov-type goodness-of-fit test.

Our trees have a special vertex called root and evolve forward in time in discrete generations; each parent node (or vertex) has up to m offspring nodes in the next generation. The set of possible vertices is called $\widetilde{V}$. A tree is a function $x : \widetilde{V} \to \{0,1\}$, where $x(v)$ indicates if the vertex $v \in \widetilde{V}$ is present in x, with the restriction that a vertex cannot be present if its mother is not. Call $\mathcal{T}$ the resulting space of trees; when $\widetilde{V}$ is finite, $\mathcal{T}$ is also finite. In the general case, $\mathcal{T}$ is a closed subset of the compact product space $\{0,1\}^{\widetilde{V}}$. Since the product topology is the one where convergence is in each coordinate, the topology may be induced by different distances. In this setting $\mathcal{B}$, the Borel σ-field, is the same as the one generated by the projections. A similar setup was proposed by Otter (1949) and Neveu (1986), see Kurata and Minami (2004). For a probability measure ν on $\mathcal{T}$, a random tree with law ν and a distance d on $\mathcal{T}$, the d-mean related to ν is defined as the tree (or set of trees) that minimizes the ν-average d-distance to the random tree. Other tree spaces and metrics are briefly discussed in Section 7.

We consider a sample of independent and identically distributed random elements of a compact metric space with law ν and a unique d-mean. We prove that the empiric d-mean of the sample converges to the d-mean related to ν as the size of the sample goes to infinity. Hence the empiric d-mean is a consistent estimator for the d-mean related to ν. The result applies to metric spaces of trees that may have infinitely many vertices. The law of large numbers on metric spaces with negative curvature has been addressed by Herer (1992), de Fitte (1997) and Es-Sahib and Heinich (1999). Our space is not of negative curvature, as shown in Section 7. For compact metric spaces, a strong law of large numbers have been obtained in Sverdrup-Thygenson (1981), which is used in our setting.

We show an invariance principle for the random processes $(g_n(y) - g(y),\ y \in \mathcal{T})$, where $g_n(y)$ is the average of the distances from y to the points of a sample of size n and $g(y)$ is the average of the distances from y to the random tree with law ν from where the sample is obtained. The proof is based on a theorem by Ledoux and Talagrand (1991); we build up a probability measure on the space of trees that satisfies the "majorizing measure condition" for a particular family of distances.

The invariance principle implies the approximate distribution of

$$\max_{y \in \mathcal{T}} |g_n(y) - g(y)|, \tag{1.1}$$

is known. We propose (1.1) as statistic for a universal Kolmogorov-type goodness of fit

test and the analogous for the two-sample problem. In general $(g_n(y) - g(y),\ y \in \mathcal{T})$ does not identify the measure ν. Busch et al (2006) show that $(g_n(y) - g(y),\ y \in \mathcal{T})$ identifies the vertex-marginals $(\nu\{x : x(v) = 1\},\ v \in \widetilde{V})$ and viceversa. The vertex-marginals do not always identify the measure but they do if the tree is constructed in a Markovian way; examples include Galton-Watson and other related processes.

As far as we know the Otter-Neveu set-up has not been used before to construct statistical tools for random trees. With this structure the law of large numbers and invariance principles are quite straightforward and the statistic (1.1) arises naturally to perform goodness-to-fit tests. The computation of the statistic (1.1) requires in principle an exponential number of steps in the number of possible vertices. Busch et al (2006) show that the search of the maximum in (1.1) is equivalent to the search of the minimal cut in an associated network; a technique coming from image reconstruction. This makes the test viable for reasonable big trees.

The critical values related to the statistic (1.1) depend on the distribution ν. To compute them it is usually necessary to simulate trees with the tested distribution or to perform bootstrap. Our test has been applied to samples of Galton-Watson related processes obtained by simulation and to a classification of FGF protein families (Busch et al 2006). In both cases the test has been successfull to distinguish different laws, even when the mean tree is the same for the two samples.

In Section 2 we introduce the space of trees as a metric space and define the d-mean tree. In Section 3 we prove the law of large numbers. In Section 4 we give some examples and in Section 5 we prove the invariance principle. In Section 6 we describe the statistical applications. In Section 7 we show that our space is not of negative curvature and discuss some other possible metrics.

2 A metric space of rooted trees

Let $\widetilde{V} = \{1, 11, 12, \ldots, 1m, \ldots\}$ the set of finite sequences of numbers in $A = \{1, \ldots, m\}$ starting with 1, with m a natural number. Elements of $\widetilde{V}$ are called *vertices*; the vertex 1 is called *root*. The *full tree* is the oriented graph $\tilde{x} = (\widetilde{V}, \widetilde{E})$ with edges $\widetilde{E} \subset \widetilde{V} \times \widetilde{V}$ given by $\widetilde{E} = \{(v, va) : v \in \widetilde{V},\ a \in A\}$, where va is the sequence obtained by juxtaposition of v and a. In the full tree each node or vertex has exactly m outgoing edges to her offsprings and one ingoing edge from her mother, except for the root that has no ingoing edges. The node $v = a_1 \ldots a_k$ is said to belong to the *generation* k; in this case we write $\mathrm{gen}(v) = k$. Generation 1 has only one node: the root of the tree.

We define a tree as a function $x : \widetilde{V} \to \{0, 1\}$ satisfying, for all $v \in \widetilde{V}$ and $a \in A$,

$$x(v) \quad \geq \quad x(va). \tag{2.1}$$

Abusing notation, we identify x with the graph $x = (V_x, E_x)$ where

$$V_x = \{v \in \widetilde{V} \, : \, x(v) = 1\}, \tag{2.2}$$
$$E_x = \{(v, va) \in \widetilde{E} \, : \, x(v) = x(va) = 1\} \, . \tag{2.3}$$

Let $\mathcal{T}$ be the set of trees of this form. Condition (2.1) in effect requires that for $x \in \mathcal{T}$, every node in x must have a parent node in each previous generation back to the root.

A finite tree is characterized by the set of its terminal nodes. For example, the trees in Figure 1 are (a) {111,12}, and (b) {11,121}.

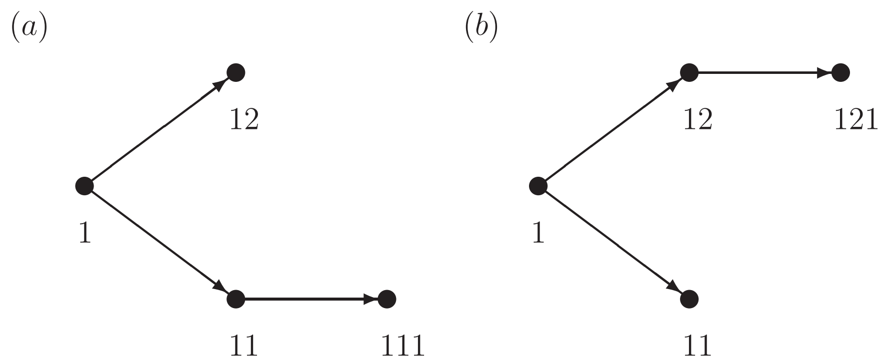

Figure 1: Two finite trees both with 3 generations and 2 terminal nodes.

The product topology on $\{0,1\}^{\widetilde{V}}$ is the smaller for which the projections are continuous. By projection we mean the family of functions $\pi_v : \{0,1\}^{\widetilde{V}} \to \{0,1\}$ that map $x \to x(v)$. In this topology x_n converges to x if and only if $x_n(v)$ converges to $x(v)$ for all $v \in \widetilde{V}$. Since at each vertex we have values in $\{0,1\}$, convergence means that for each v there exists $n(v)$ such that if $n \geq n(v)$ then $x_n(v) = x(v)$. This condition guarantees that $\mathcal{T}$, the space of trees, is a closed set in $\{0,1\}^{\widetilde{V}}$ and hence also compact.

As it is done in interacting particle systems (see Liggett (1985)) we consider the sigma algebra $\mathcal{B}$ generated by the cylinders $\{x \in \mathcal{T} \, : \, x(v) = 1\}$, $v \in \widetilde{V}$; this is just the Borel sigma field generated by the product topology.

We provide $\mathcal{T}$ with a *distance* d, so that $(\mathcal{T}, d)$ is a *metric space*. We use the family of distances in $\mathcal{T}$ defined by

$$d(x,y) = \sum_{v \in \widetilde{V}} |x(v) - y(v)| \phi(v) \, , \tag{2.4}$$

for some strictly positive function $\phi : \widetilde{V} \to \mathbb{R}^+$ satisfying $\sum_{v \in \widetilde{V}} \phi(v) < \infty$. In this case, the distance between the two trees of Figure 1 is $d(a,b) = \phi(111) + \phi(121)$.

This distance is compatible with the product topology and hence, the notion of convergence under any of these metrics is the same as the induced by the product topology.

Otter (1949) and Neveu (1986) propose a similar construction, but to deal with unbounded number of offsprings, they ask each vertex v and natural a to satisfy $x(v(a+1)) \leq x(va)$; informally, the presence of a brother in the tree implies that all older brothers are also present. Their distance, also compatible with the product topology, is defined by

$$d_{ON}(x,t) = \exp\big(- \max\{k \,:\, x(v) = t(v) \text{ for all } v \text{ such that } \mathrm{gen}(v) \leq k\}\big). \tag{2.5}$$

See Kurata and Minami (2004) for a review of those papers.

Random trees A *random tree* with distribution ν is a measurable function

$$T : \Omega \to \mathcal{T} \text{ such that } \mathbb{P}(T \in A) = \int_A \nu(dx)\ . \tag{2.6}$$

for any Borel set $A \in \mathcal{B}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and ν a probability on $(\mathcal{T}, \mathcal{B})$.

The expected distance from a tree y to a random tree T is defined by

$$g(y) \;:=\; \mathbb{E}(d(T,y)) \;=\; \int_{\mathcal{T}} d(x,y)\, \nu(dx) \tag{2.7}$$

$$= \sum_{x \in \mathcal{T}} \nu(x) d(x,y) \quad \text{(in the discrete case).} \tag{2.8}$$

Definition 2.1 *The expected value or d-mean of a random tree T is the set (of trees) $\mathbb{E}_d T$ that minimizes the expected distance to T:*

$$\mathbb{E}_d T \;:=\; \operatorname*{arg\,min}_{y \in \mathcal{T}} g(y)\,. \tag{2.9}$$

The set $\mathbb{E}_d T$ might be empty, but if $\mathcal{T}$ is compact, then $\mathbb{E}_d T$ is not empty (see Section 3). Any element of the set $\mathbb{E}_d T$ is also called a d-mean. Since $\mathbb{E}_d T$ depends only on the distribution ν induced by T on $\mathcal{T}$, it may also be denoted as $\mathbb{E}_d(\nu)$. The elements of $\mathbb{E}_d(\nu)$ are also called d-centers. The notion of expected value depends on the distance d; in particular, for random variables in $\mathbb{R}^k$ we may obtain the usual mean, the median and the mode as illustrated in Section 4.

Example: In the Galton-Watson branching process the numbers of offspring of distinct nodes are i.i.d. In the special case that they have the Binomial(2,p) distribution with $p \in [0,1]$, the offspring number is 0, 1, or 2 with probabilities $(1-p)^2$, $2p(1-p)$, and p^2. Letting $k_0 = \max\{k \in \{0,1,\dots\} : p^k \geq 1/2\}$ there are two cases: (a) if $p^{k_0} > 1/2$ there is only one mean tree x satisfying $x(v) = 1$ if and only if $\mathrm{gen}(v) < k_0$ and (b) if $p^{k_0} = 1/2$ the mean tree is the set of trees with $x(v) = 1$ if $\mathrm{gen}(v) < k_0$, $x(v) \in \{0,1\}$ if $\mathrm{gen}(v) = k_0$ and $x(v) = 0$ if $\mathrm{gen}(v) > k_0$. In particular, if $p < 1/2$ the mean tree is the empty tree.

Let $T_1, \ldots, T_n$ be a random sample of T (independent random trees with the same law as T). The empiric measure associated to the sample is denoted by μ_n and it is given by

$$\mu_n \ := \ \frac{1}{n}\sum_{i=1}^{n} \delta_{T_i} \ , \quad \mu_n(x) \ := \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i = x\}} \ , \tag{2.10}$$

where δ_x is the point mass at x and $\mathbf{1}_A$ is the indicator function of the set A. Associated to this measures we define the empiric expected distance of a tree y to the sample by

$$g_n(y) := \int_{\mathcal{T}} d(x,y)\ \mu_n(dx) = \frac{1}{n}\sum_{i=1}^{n} d(T_i, y)\,, \tag{2.11}$$

and as in (2.9) the *empiric mean* tree (empiric d-center, sample d-mean) as the random set given by

$$\overline{T}_n := \arg\min_{y\in\mathcal{T}} g_n(y)\,. \tag{2.12}$$

The empirical mean is like a consensus tree. If n is odd, the empirical mean is unique; it just includes all vertices that are in more than half of the trees. If n is even, it is not unique but there is a "shortest" and "largest" empirical mean tree, and every subtree of the largest empirical mean tree which contains the shortest empirical mean tree is on the set of empirical d-means. This is a nice property from the robustness point of view.

3 Law of large numbers

If ν is defined on a finite set of trees the following law of large numbers follows immediately.

Theorem 3.1 *Let $(\mathcal{T}_0, d)$ be a finite tree space with metric d. Let $T \in \mathcal{T}_0$ be a random tree with law ν such that $\mathbb{E}_d T$ has only one element (also denoted by $\mathbb{E}_d T$). Let $\{T_n, \, n \geq 1\}$ be an i.i.d. sequence of random trees with law ν. If y_n is any of the empiric mean trees of $\{T_1, \ldots, T_n\}$, that is $y_n \in \overline{T}_n$, then*

$$\lim_{n\to\infty} d(y_n, \mathbb{E}_d T) = 0 \qquad a.s.\,. \tag{3.1}$$

In other words, the set of empiric mean trees coincides with the singleton of the d-mean if n is large enough.

When ν is an arbitrary probability measure on $\mathcal{T}$, it may give positive mass to sets of trees with infinitely many nodes. First we state the strong law of large numbers for random elements taking values in a compact metric space given in Sverdrup-Thygeson (1981). This covers the space of trees with infinite number of vertices. Then we show that the metric

space $\mathcal{T}$ is compact; this implies in particular that the expected tree is well defined ($\mathbb{E}_d(T)$ is non empty).

Consider a compact metric space $(\mathcal{K}, d)$. Let $\mathcal{B}$ denote the σ-field generated by the open sets, and so the elements of $\mathcal{B}$ are the Borel sets. Let ν be a probability measure on $\mathcal{B}$. We define the expected value with respect to the measure ν and the distance d following the ideas developed in the previous section. Let $g\colon \mathcal{K} \to \mathbb{R}^+$ be given by

$$g(y) := \int_{\mathcal{K}} d(y, x)\, \nu(dx). \tag{3.2}$$

Since

$$|g(y) - g(t)| \leq \int_{\mathcal{K}} |d(y, x) - d(t, x)|\, \nu(dx) \leq \int_{\mathcal{K}} d(y, t)\nu(dx) = d(y, t)\ , \tag{3.3}$$

we get that g is Lipschitz continuous. Since it is defined on a compact space, it attains its minimum. This shows that the d-mean set $\mathbb{E}_d(\nu)$ defined as in (2.9) is non empty. The empiric mean $\overline{T}_n$ is defined as in (2.12).

Theorem 3.2 (Sverdrup-Thygeson, 1981) *Let ν be a probability on the compact metric space $(\mathcal{K}, d)$ such that $\mathbb{E}_d(\nu)$ has only one point. Consider $\{T_n : n \geq 1\}$, an i.i.d. sample for ν. Then, the empirical d-centers converge uniformly to $\mathbb{E}_d(\nu)$ almost surely:*

$$\lim_{n\to\infty} \sup_{a \in \overline{T}_n} d\Big(a, \mathbb{E}_d(\nu)\Big) = 0 \quad a.s.\,. \tag{3.4}$$

The results of this section can be extended to the following family of functions g_p defined for $p \geq 1$ by

$$g_p(y) \ = \ \int_{\mathcal{K}} d(y, x)^p\, \nu(dx)\ .$$

4 Examples

Mode parameter Consider a finite space $\mathcal{K}$ with the discrete distance given by

$$d(x, y) \ = \ \begin{cases} 0 & \text{if } x = y\ , \\ 1 & \text{otherwise.} \end{cases} \tag{4.1}$$

In this case,

$$g(x) = \int_{\mathcal{K}} d(x, y)\, \nu(dy) = \sum_{y \neq x} \nu(y) = 1 - \nu(x)\ . \tag{4.2}$$

So, the d-center parameter for $(\mathcal{K}, d, \nu)$ is just the mode of ν.

Mean and median parameters Consider $\mathcal{K} = [0,1]^n \subset \mathbb{R}^n$, and $d(x,y) = ||x-y||^p$. Let ν be any probability measure on $\mathcal{K}$. Then, if $p=2$ we have that the d-center parameter is the usual expected value. For $n=1$ and $p=1$ we get the median, and for $n>1$ the spatial median or multivariate $L_1-median$, see for instance Haldane (1948) and Milasevic and Ducharme (1987).

Product Space We say that $(\mathcal{K}, d, \nu)$ is a centered space if it has a unique d-center. We now prove that the product of a finite number of centered spaces is a centered space.

Lemma 4.1 *Let $(\mathcal{K}_i, d_i, \nu_i)$ be spaces with unique d-centers $C_i = E_{d_i}(\nu_i)$, for $i = 1,2$. Then, if we consider the product space $\mathcal{K} = \mathcal{K}_1 \times \mathcal{K}_2$ with*

$$d(\hat{x}, \hat{y}) = d_1(x_1, y_1) + d_2(x_2, y_2)\,, \tag{4.3}$$

for $\hat{x} = (x_1, x_2) \in \mathcal{K}$ and the product measures $\nu = \nu_1 \times \nu_2$, we get that $(\mathcal{K}, d, \nu)$ has also a unique d-center (C_1, C_2).

Proof We need to prove that (C_1, C_2) is the unique point minimizing $g : \mathcal{K} \to \mathbb{R}$. We get

$$g(\hat{x}) = \int_{\mathcal{K}_1} \int_{\mathcal{K}_2} \Big(d_1(x_1, y_1) + d_2(x_2, y_2) \Big) \nu_2(dy_2) \nu_1(dy_1) = g_1(x_1) + g_2(x_2)\ , \tag{4.4}$$

where

$$g_i(x) = \int_{\mathcal{K}_i} d_i(x, y) \nu_i(dy)\,. \tag{4.5}$$

from where the result follows. □

5 Invariance Principle

In this section we consider a sequence of independent identically distributed random trees $(T_1, \ldots, T_n)$ with empiric mean $g_n(t)$ given by (2.11) and prove an invariance principle for the centered process

$$(\sqrt{n}(g_n(t) - g(t)), t \in \mathcal{T})\,,$$

as $n \to \infty$. The main tool is the following general result.

Theorem 5.1 (Ledoux and Talagrand (1991) pag 395–396) *Let $\mathcal{T}$ be a compact metric space and $C(\mathcal{T})$ be the separable Banach space of continuous functions on $\mathcal{T}$ with the sup norm. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to C(\mathcal{T})$ be a random element of $C(\mathcal{T})$ with $\mathbb{E}X(t) = 0$ and $\mathbb{E}X(t)^2 < \infty$ for all t in $\mathcal{T}$. Assume that X is Lipschitz, that is, there exists a positive random variable M with $\mathbb{E}M^2 < \infty$ such that*

$$|X(\omega, s) - X(\omega, t)| \;\leq\; M(\omega)\, d(s,t)\,, \tag{5.1}$$

for all $\omega \in \Omega$, $s, t \in \mathcal{T}$. Assume there exists a probability measure μ on $(\mathcal{T}, d)$ such that

$$\lim_{\delta\to 0} \sup_{t\in\mathcal{T}} \int_0^\delta \Big[-\log[\mu(B(t,u))]\Big]^{1/2} du \;=\; 0\,, \tag{5.2}$$

where $B(t, u)$ is the ball centered at t with radius u. (This is called the majorizing measure condition for $(\mathcal{T}, d)$.) Then X verifies the Central Limit Theorem in $C(\mathcal{T})$. That is, if $X_1, \ldots, X_n$ are i.i.d. with the same law as X, then $n^{-1/2}(X_1 + \ldots + X_n)$ converges to a Gaussian process with mean zero and the same covariance function as X.

The majorizing measure condition If $\mathcal{T}$ is finite, the condition is satisfied automatically by any measure μ on $\mathcal{T}$ giving positive mass to all elements of $\mathcal{T}$. Indeed, $\mu(B(t, u)) \geq \mu(t) > 0$ and the integral in (5.2) is dominated by $[-\log(\mu(t))]^{1/2}\delta$.

Lemma 5.2 *Let $\mathcal{T}$ be the set of trees. Let $0 < z < m^{-3/2}$ and ϕ defined by*

$$\phi(v) = z^{\mathrm{gen}(v)}\,. \tag{5.3}$$

Then the majorizing measure condition is satisfied for $(\mathcal{T}, d)$ with the distance defined by (2.4) and this ϕ.

Proof Since for finite trees the result follows, we assume the trees in $\mathcal{T}$ have infinitely many generations. Define the cylinder of generation k induced by the tree $t \in \mathcal{T}$ by

$$\mathcal{T}_k(t) := \{s \in \mathcal{T} \,:\, s(v) = t(v) \text{ if } \mathrm{gen}(v) \leq k\}\,. \tag{5.4}$$

Define for $u \geq 0$

$$k(u) \;=\; k(u, \phi) \;\;:=\;\; \inf\Big\{k \,:\, \sum_v \phi(v)\mathbf{1}\{\mathrm{gen}(v) > k\} < u\Big\}\,.$$

Since $\sum_v \phi(v) < \infty$, $k(u)$ goes to ∞ as u goes to 0. Since

$$\sum_v \phi(v)\mathbf{1}\{\mathrm{gen}(v) > k\} = \sum_{i>k} m^{i-1}z^i = \frac{z(mz)^k}{1-mz}\,, \tag{5.5}$$

we can write

$$k(u) = \inf\{k \,:\, z(mz)^k/(1-mz) < u\}\,. \tag{5.6}$$

We have

$$\mathcal{T}_{k(u)}(t) \subset B(t, u)\,. \tag{5.7}$$

A natural choice for a majorizing measure in $\mathcal{T}$ is the measure induced by the product measure ν_ρ on $\{0, 1\}^{\widetilde{V}}$ with marginals $\nu_\rho\{\xi \,:\, \xi(v) = 1\} = \rho$, for $v \in \widetilde{V}$. Given a configuration $\xi \in \{0, 1\}^{\widetilde{V}}$, define $x(\xi)$ as the maximal tree from the root whose vertices are contained in the set ξ. In other words, inductively, $x(\xi)(1) = \xi(1)$ and

$$x(\xi)(va) \ := \ \begin{cases} 1 & \text{if } x(\xi)(v)=1 \text{ and } \xi(va)=1 \\ 0 & \text{otherwise ,} \end{cases} \tag{5.8}$$

for each $v \in \{0,1\}^{\widetilde{V}}$ and $a \in A$. Define the measure μ_ρ induced on $\mathcal{T}$ by this application:

$$\mu_\rho(B) := \nu_\rho\{\xi : x(\xi) \in B\} \ .$$

To check that μ_ρ is a majorizing measure, let $\beta > 0$ be defined by $e^{-\beta} = \min\{\rho, 1-\rho\}$. The number of vertices in the first k generations of the full tree is $(m^k - 1)/(m-1) \leq 2m^k$. Hence the probability of any cylinder with k generations is bigger than $e^{-2\beta m^k}$:

$$\mu_\rho(\mathcal{T}_k(t)) \geq \nu_\rho\Big\{\xi \in \{0,1\}^{\widetilde{V}} : \xi(v) = t(v) \text{ if } \mathrm{gen}(v) \leq k\Big\} \geq e^{-2\beta m^k} \ . \tag{5.9}$$

uniformly in t. This and (5.7) imply that the supremum of the integral in (5.2) is bounded above by

$$\int_0^\delta (2\beta)^{1/2} m^{k(u)/2} du \ = \ \int_0^\delta (2\beta)^{1/2} e^{k(u)\log(m^{1/2})} du \ \leq \ (2\beta)^{1/2} \int_0^\delta \frac{1}{u^{1-\varepsilon}} du \, , \tag{5.10}$$

for δ small enough, if there exists an $\varepsilon > 0$ such that $k(u) \leq -(\log u)(1-\varepsilon)/\log(m^{1/2})$, for u small enough. In this case the proof is finished because for $\varepsilon > 0$ (5.10) converges to zero as $\delta \to 0$. Call $\gamma = (1-\epsilon)/\log(m^{1/2})$. In view of (5.5), we look for $\gamma > 0$ such that $(mz)^{-\gamma \log u} < u(1-mz)/z$. That is,

$$u^{-\gamma \log(mz) - 1} < \frac{1-mz}{z} \ .$$

For u sufficiently small it suffices that $-\gamma\log(mz) - 1 > 0$ and $z < m^{-1}$. Substituting γ and noticing that $\log(mz) < 0$, we need to find an $\varepsilon > 0$ such that

$$-(1-\varepsilon) < \frac{\log(m^{1/2})}{\log(mz)}, \qquad \text{that is,} \qquad \varepsilon < 1 + \frac{\log(m^{1/2})}{\log(mz)} \ ,$$

which exists since $z < m^{-3/2}$. $\square$

We are now able to obtain the asymptotic distribution of the process

$$\sqrt{n}\,(g_n(t) - g(t)) = \frac{\sum_{i=1}^n [d(T_i,t) - \mathbb{E}(d(T_i,t))]}{\sqrt{n}}.$$

Theorem 5.3 *Let $\mathcal{T}$ be the set of trees with at most m offspring. Consider the distance given in (2.4) for $\phi(v) = z^{\mathrm{gen}(v)}$ with $0 < z < m^{-3/2}$. Let $\{T_i : i \geq 1\}$ be a sequence of i.i.d. random trees on $\mathcal{T}$ with the same law as T. Then the process $(\sqrt{n}(g_n(t) - g(t)), t \in \mathcal{T})$ converges weakly as $n \to \infty$ to a Gaussian process W with zero mean and the same covariance function as the process $X \in (\mathbb{R}^+)^{\mathcal{T}}$ defined by $X(t) = d(T,t) - \mathbb{E}(d(T,t))$.*

Proof Since $|\,X(\omega,t) - X(\omega,t')| \leq 2d(t,t')$ the result follows from the previous Lemma and Theorem (5.1). $\square$

6 Statistical applications

Let T be a random tree in $\mathcal{T}$ with distribution ν and mean distances $(g(y),\, y \in \mathcal{T})$ defined in (2.7). Let ν_0 be a distribution on the tree space $\mathcal{T}$ with mean distances $(g_0(y),\, y \in \mathcal{T})$. The goal is to test

H0: $\nu = \nu_0$
HA: $\nu \neq \nu_0$

using an i.i.d. sample of random trees $\{T_i : i \geq 1\}$. Notice however that the rejection of H0 does not imply the rejection of $\mathbb{E}T = \mathbb{E}_d(\nu_0)$.

To perform the test we propose the statistic

$$\sup_{y\in\mathcal{T}} |W_n(y)| = \sup_{y\in\mathcal{T}} \sqrt{n}\,|\, g_n(y) - g_0(y))\,|, \tag{6.1}$$

whose asymptotic law under H0 is obtained from Theorem (5.3) and the Continuous Mapping Theorem. We reject the null hypothesis at level α if

$$\sup_{y\in\mathcal{T}} |W_n(y)| > q_\alpha\ ,$$

where q_α satisfies $P(\sup_{y\in\mathcal{T}} |W(y)| > q_\alpha) = \alpha$, for W given in Theorem (5.3) under $\nu = \nu_0$.

The test rejects $\nu = \nu_0$ if g determines ν unequivocally.

In practice the distribution of $\sup_{y\in\mathcal{T}} |W(y)|$ depends on the covariance of the process $X(t) = d(T,t) - \mathbb{E}(d(T,t))$ which in general is unknown. A possible way to deal with this problem is to approximate q_α using bootstrap. The validity of the bootstrap in this context remains an open problem. Alternatively, one can simulate trees with distribution ν_0 and estimate q_α.

For the problem of two samples (of same size, for instance) one may use the statistic

$$\sup_{y\in\mathcal{T}} \sqrt{n}|g_n(y) - g'_n(y)|\,, \tag{6.2}$$

where g_n and g'_n correspond to the samples of T and T' respectively.

When g characterizes the measure ν? Busch et al (2006) prove that $g = (g(t), t \in \mathcal{T})$ characterizes the vertex-marginal distributions as follows. Let ν and ν' be two measures in $\mathcal{T}$ and g, g' be the corresponding processes. Then $g = g'$ if and only if $\nu\{t : t(v) = 1\} = \nu'\{t : t(v) = 1\}$ for all vertex v. In that paper it is proven that under certain Markov hypothesis, the vertex-marginals identify univoquely the measure. The class of random trees satisfying those hypothesis includes Galton-Watson branching processes and other related processes.

7 Metrics and negative curvature

In this section we show that our tree space cannot be embedded in a metric space of non positive curvature. Then we discuss other possible metrics that have been considered for spaces of trees. A natural way of embedding the discrete tree space $\mathcal{T}$ in a continuous space would be to consider a tree as a function $x : \widetilde{V} \to \mathbb{R}^+$ (instead of $\{0,1\}$), where the value $x(v)$ would represent the length of the edge connecting the node v to her mother. The value $x(v) = 0$ means that the node v is not present. The metric could be the one given in (2.4) which coincides with the previous one for trees with unitary edge lengths. A tree condition like "$x(va) > 0$ implies $x(v) > 0$" is also needed, but other conditions could be proposed. For instance one could collapse the vertices with $x(v) = 0$ but in this case the trees would not have a limited number of offspring nor the vertex notation introduced in Section 1 would be appropriate.

Let a, b, c, x be arbitrary distinct points in a metric space $\mathcal{T}$ such that x belongs to a geodesic from a to b, that is, $d(a,b) = d(a,x) + d(x,b)$. Let a', b', c', x' in $\mathbb{R}^2$ be points located in such a way that the relative distances are the same, that is, $d(a,b) = d'(a',b')$, $d(a,c) = d'(a',c')$, etc, where d' is the Euclidean distance in $\mathbb{R}^2$. It is said that $\mathcal{T}$ is of *non positive curvature* if $d(x,c) \leq d'(x',c')$ for any choice of a, b, c, x. These spaces are also called $CAT(0)$, see Billera, Holmes and Vogtmann (2001), page 750.

We now give an example showing that our space cannot fit the above property. Let $a = \{111, 12\}$ and $b = \{11, 121\}$ be the trees in Figure 1 and $x = \{11, 12\}$ and $c = \{111, 112, 121, 122\}$ those of Figure 2. Consider the distance (2.4) with $\phi(v)$ depending only on the generation of v, so that $\phi(111) = \phi(121) = \phi(122) = \phi(112) = \alpha$, for some $\alpha > 0$. The tree x belongs to a geodesic between a and b: $d(a,x) = d(b,x) = \alpha$, $d(a,b) = 2\alpha$. On the other hand $d(a,c) = d(b,c) = 3\alpha$ and $d(x,c) = 4\alpha$. Consider the corresponding Euclidean triangle (a', b', c') with the same relative distances. The point equidistant from a' and b' in the Euclidian geodesic, corresponding to x, is $x' = (a' + b')/2$. Since $d'(x',c') = \sqrt{8}\alpha < 4\alpha = d(x,c)$, our tree space cannot be embedded in a $CAT(0)$ space.

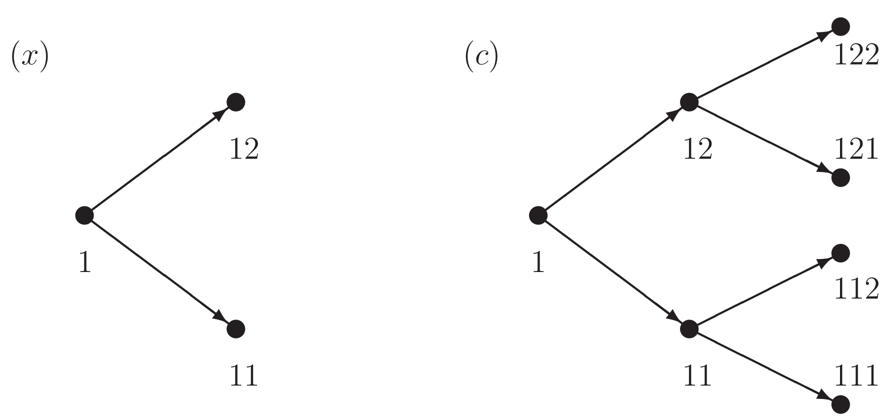

Figure 2: The trees $x = \{11, 12\}$ and c={111,112,121,122}.

Notice that the tree $\bar{x} := (111, 121)$ is also in a (different) geodesic between a and b. However $d(\bar{x}, c) = 2\alpha < \sqrt{8}\alpha = d(x, c)$. In fact the triangle (a', b', c') is the same and $\bar{x}' = x'$. Actually, the fact that there are two geodesics going from a to b indicates that the space cannot be of negative curvature.

The metric we propose for the space of trees is usual in interacting particle systems, some of which are defined in $\{0, 1\}^S$ for countable S, for example. The product of the discrete topologies induces a metric like (2.4). Under this metric, the convergence of a sequence x_n to x is equivalent to the convergence of $x_n(v)$ to $x(v)$ for all vertex v. Valiente (2001) considers spaces of finite trees with ordered vertices, reviews several distances and proposes a new metric. An important example is the so called "edit distance", which counts the number of operations (eliminate a vertex, add a vertex) that need to be done in order to transform one tree into another one. Critchlow (1980) proposes some metrics in the set of permutations of a finite sequence that may be adapted to a finite space of trees. It would be nice to understand if our results can be proven in those spaces.

Billera, Holmes and Vogtmann (2001) describe various spaces of "phylogenetic trees" and construct a (continuous) convex metric space of trees with a fixed number n of final vertices (i.e., vertices with no daughters). The resulting space $\mathcal{F}_n$ is $CAT(0)$. Phylogenetic trees are constructed from the final vertices to the root by successively grouping subsets of vertices as in the Kingman's coalescent. Each vertex with descendants represents the most recent common ancestor of the descendants, and the length of the edge (v, v') represent the time a group of species represented by v' needed to split. In our space the trees can have variable number of final vertices; our counterexample does not apply to spaces of trees with fixed number of final vertices. Another difference with phylogenetic trees is that in our space we do not label the (final) vertices.

8 Final remarks

Our motivation was to produce a statistical tool to study the asymptotic behavior of sequences of random trees. The law of large numbers is not directly applied to construct the tests, but is important to guarantee the consistency of the estimators. On the other hand, the central limit theorem (Theorem 5.3) uses the tree structure and a particular form of the distance. The shape of the function ϕ intervening in the distance was necessary to show that the majorizing measure condition holds (Lemma 5.2). We believe this can be extended to other structures contained in a subset of $\{0, 1\}^S$ for S countable. Another possible extension is to eliminate the upperbound m on the number of offsprings. If the mean number of offsprings is not finite, then the limits may be stable laws, but this is to be stablished.

The statistical application we have considered in Section 7 points in the direction of a Kolmogorov-Smirnov type goodness of fit test. We are interested in the decision problem: given a random sample $T_1, ..., T_n$ can we decide if their underlying common distribution P is

a given P_0? For instance does the sample follow the Galton-Watson model with parameter p_0? We think that the statistic given in section 7 is adequate for this problem. The results in Busch et al (2006) where our test has been applied to several simulated examples, and a real data example to classify FGF protein families points in this direction.

The implementation of the tests requires the computation of the statistic (6.1) which is a supremum over the space of trees of the distance of the tree to the mean tree. The computation time of this task may increase fast with the number of nodes. Busch et al (2006) propose a method to transform this problem in the computation of the minimal cut of the flux of a related graph. This allows to see the behavior of the test in some concrete examples related to Galton-Watson generated random trees.

Acknowledgements

We thank one of the referees for mentioning that our setup was proposed by Otter and Neveu. We thank all referees for their careful reading and constructive comments.

This work started in the open problem session of the 6th Brazilian School of Probability, Ubatuba, 2002. This paper is partially supported by the Institute of Millenium for the Global Advance of Mathematics, Fundação de Amparo à Pesquisa do Estado de São Paulo, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Programa Núcleos de Excelência and the Brazil-Argentina Agreement CAPES-SECyT.

References

[1] Billera, L. J.; Holmes, S. P.; Vogtmann, K. (2001) Geometry of the Space of Phylogenetic Trees, *Advances in Applied Mathematics,* **27**, 4:733-767.

[2] Busch, J. R.; Ferrari, P. A.; Flesia, A. G.; Fraiman, R.; Grynberg, S. ; Leonardi, F. (2006) Testing statistical hypothesis on random trees. Preprint.

[3] De Fitte, P. R. (1997) Théorème ergodique ponctuel et lois fortes des grandes nombres pour des points aléatoires d'un espace métrique à courbure négative. *Ann. Probab.* **25** 2:738-766.

[4] Es-Sahib, A.; Heinich, H. (1999) Barycentre canonique pour un espace métrique à courbure négative. *Séminaire de Probabilités, XXXIII, 355–370, Lecture Notes in Math.*, **1709**, Springer, Berlin.

[5] Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35**, 414-415.

[6] Herer, W. (1992) Mathematical expectation and strong law of large numbers for random variables with values in a metric space of negative curvature. *Probab. Math. Statist.* **13** 1:59-70.

[7] Kingman, J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.,* **13** 235-248.

[8] Kurata, K. and Minami, N. (2004). The equivalence of two constructions of Galton–Watson processes. Preprint mp_arc 04-295.

[9] Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces. Isoperimetry and Processes.* Springer-Verlag.

[10] Liggett, T. M., (1985) *Interacting Particle Systems*. Springer-Verlag, New York.

[11] Milasevic, P. and Ducharme, G. R. (1987). Uniqueness of the spatial median. *Annals of Statistics* **15**, 1332-1333.

[12] Mourier, E. (1953) Eléments aléatoires dans un espace de Banach. *Ann. Inst. Henri Poincaré,* **13** 159-244.

[13] Neveu, J. (1986). Arbres et processus de Galton-Watson. *Ann. Inst. H. Poincaré, Probabilités et Statistique* **22,** 199–207.

[14] Otter, R. (1949). The Multiplicative Process. *Ann. Math. Stat.* **20,** 206–224.

[15] Sverdrup-Thygeson, H. (1981). Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces. *Ann. Statist.,* **9** 141-145.

[16] Valiente, G. (2001) An Efficient Bottom-Up Distance between Trees. *Proc. 8th Int. Symposium on String Processing and Information Retrieval*, IEEE Computer Science Press. SPIRE 2001: 212-219.